Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 14

## 1. Aplicación del teorema de límite central

El teorema del límite central (CTL por sus siglas en inglés de, *Central Limit Theorem* es un teorema fundamental de probabilidad y estadística. Este teorema indica que cuando recolectamos una muestra suficientemente grande de $n$ observaciones independientes de una población con media $\mu$ y varianza finita $\sigma^2$, la distribución muestral de las medias muestrales sigue aproximadamente una distribución normal con media = $\mu$ y desviación estándar = $\sigma/\sqrt{n}$ [1].

Sean $X_1$, $X_2$, $\ldots, X_n$ un proceso de pruebas independientes e igualmente distribuidos (i.i.d) con media $\mu$ y una varianza finita $0 < \sigma^2 < \infty$ y sean $S_n = X_1 + X_2 + \cdots + X_n$ y $\Phi z$ la función de densidad de la distribución normal estándar, en la ecuación 1 se expresa una versión del teorema [5]:

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = \Phi(z). \tag{1}$$

**Principales propiedades del teorema central del límite**

De acuerdo a López [3] el teorema central del límite tiene una serie de propiedades de gran utilidad en el ámbito estadístico y probabilístico. Las principales son:

- Si el tamaño de la muestra es suficientemente grande, la distribución de las medias muestrales seguirá aproximadamente una distribución normal, lo cual se cumple independientemente de la forma de la distribución con la que se trabajando.
- En estadística, un tamaño de muestra lo suficientemente grande para sacar conclusiones es mayor o igual a 30.

- Selección al azar, para que no esté sesgado hacia ciertas características de la población.
- La media poblacional y la media muestral serán iguales. Es decir, la media de la distribución de todas las medias muestrales será igual a la media del total de la población.

El teorema del límite central es una aproximación que se puede usar cuando la población que está estudiando es tan grande que tomaría mucho tiempo recopilar datos sobre cada individuo que forma parte de ella, por ende, en términos estadísticos, al recolectar muestras de una población en particular y combinar la información de las muestras, se podría sacar conclusiones sobre la población [1]. Además, cuando los datos tienden a tener una distribución normal este tipo de distribucion es muy fácil de aplicar para realizar contrastes de hipótesis y construcción de intervalos de confianza [3].

El caso aplicado de este teorema fue con base en el trabajo de Bento [1] donde se utiliza el CTL para determinar la cantidad de reabastecimiento semanal de un producto y no incurrir en exceso de inventario inactivo en las tiendas de una cadena de supermercados. Como se mencionó anteriormente, con base en el teorema del límite central, no es necesario tener que visitar todas las tiendas de la región y obtener las cifras de ventas del producto de la semana para saber cuántas unidades solicitar en el próximo pedido. Lo que se puede hacer es recopilar muchas muestras de las ventas semanales en las tiendas (la población), calcular su media (el número medio producto vendido) y construir la distribución de las medias de la muestra, conocida como la distribución muestral. Si estas muestras cumplen los criterios del CTL, se podrá que la distribución de las medias muestrales se puede aproximar a la distribución normal. Los criterios de la muestra serían los siguientes:

- Seleccionado al azar para evitar sesgo.
- Representativa de la población, mayor o igual a 30.
- Selección al azar, para que no esté sesgado hacia ciertas características de la población.
- Incluir menos del 10 % de la población, si se toman muestras sin reemplazo ya que que las observaciones en la población no son todas independientes entre sí, si se recolecta una muestra que es demasiado grande, puede terminar recolectando observaciones que no son independientes entre sí. Incluso si esas observaciones se eligieron al azar.

Para la experimentación se utilizó una base de datos con datos de ventas, información de pedido, entre otros, de dominio público disponible en Kaggle [4]. Se realizó una simulación como ejemplo numérico para demostrar el CLT, es decir, con una muestra representativa de la población bajo los criterios mencionados anteriormente, la distribución de las medias muestrales a medida que aumenta el número de muestras tomadas, más se acerca a la forma de una distribución normal. Para esta experimentación, se calcula el promedio de ventas con 10, 100, $1,000$ y $10,000$ muestras de tamaños 30 y 100. El código

empleado para esta simulación se realizó en el programa R versión 4.0.2 [6] y junto con la base de datos utilizada se encuentran en el repositorio GitHub [2].

En la figura 1 se muestra el histograma de las ventas totales obtenidas, en la cual podemos observar que tienen una media aproximada de $5,553.89$. En la figura 2 y 3 se muestran los resultados obtenidos para las 10, 100, $1,000$ y $10,000$ muestras de tamaño 30 y 100 respectivamente, en las cuales podemos observar que con $10,000$ muestras aleatorias de tamaño 100 del conjunto de datos de ventas, se obtiene una distribución de muestreo que se asemeja a la curva de campana característica de la distribución normal con un promedio de ventas similar al de la población (ventas totales). Lo anterior sucede ya que, al recolectar una muestra más grande, se tendrá menos posibilidades de obtener valores extremos, por lo que sus valores estarán más agrupados y por lo tanto, la desviación estándar, o la distancia de la media, será menor. Otra forma de explicarlo sería considerando la ecuación 1, ya que la desviación estándar de la distribución muestral, también llamada error estándar, es igual a $\sigma/\sqrt{n}$, entonces, a medida que aumenta el tamaño de la muestra, el denominador también aumenta y hace que el valor estándar general sea más pequeño [1].

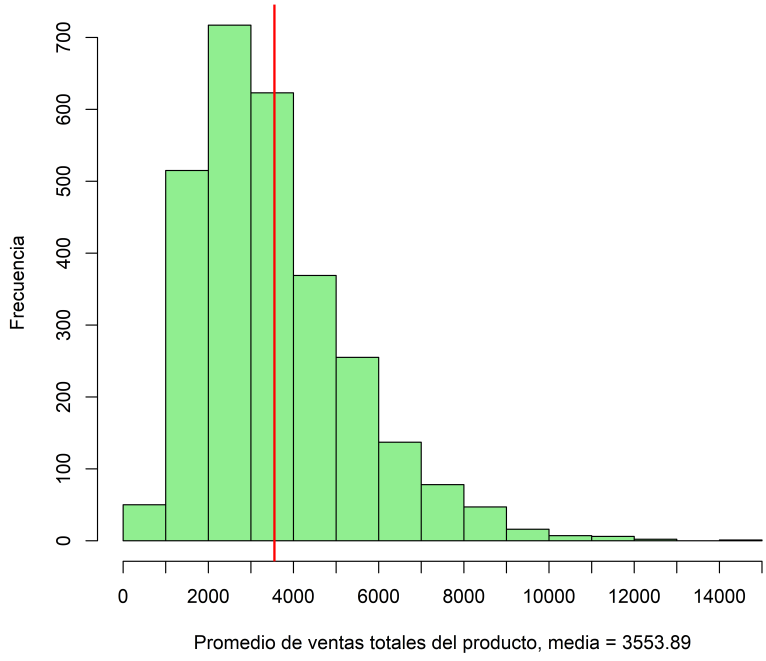



Figura 1: Histograma de las ventas totales del producto

## Referencias


[1] Bento, Carolina. Central Limit Theorem: a real-life application. Recurso disponible en, https://towardsdatascience.com/central-limit-theorem-a-real-life-application-f638657686e1, 2015.

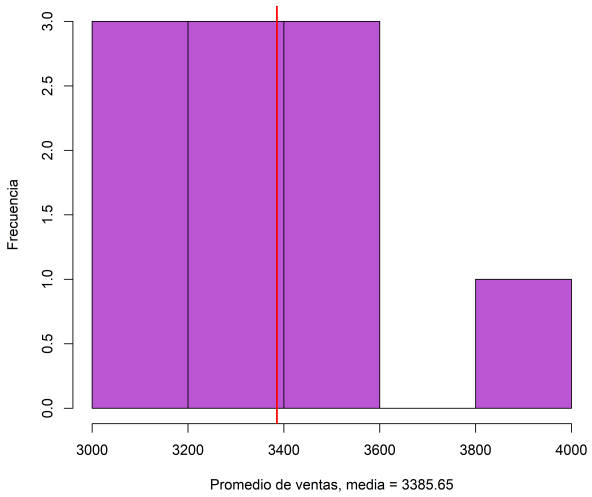
[2] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea14, 2020.

[3] López Abellán, Joaquín. Teorema central del límite (TCL). Recurso disponible en, https://economipedia.com/definiciones/teorema-central-del-limite.html, 2015.

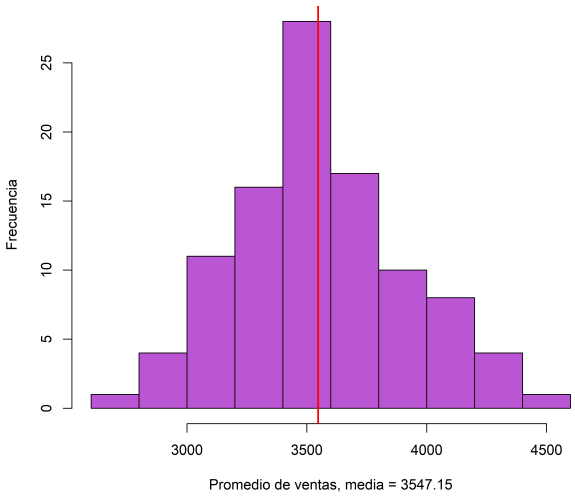
[4] Segura, Gus. Sample Sales Data. Recurso disponible en, https://www.kaggle.com/kyanyoga/sample-sales-data, 2016.

[5] Stanton, Charles. The Central Limit Theorem. Recurso disponible en, https://web.archive.org/web/20100602111757/http://www.math.csusb.edu/faculty/stanton/probstat/clt.html, 2010.

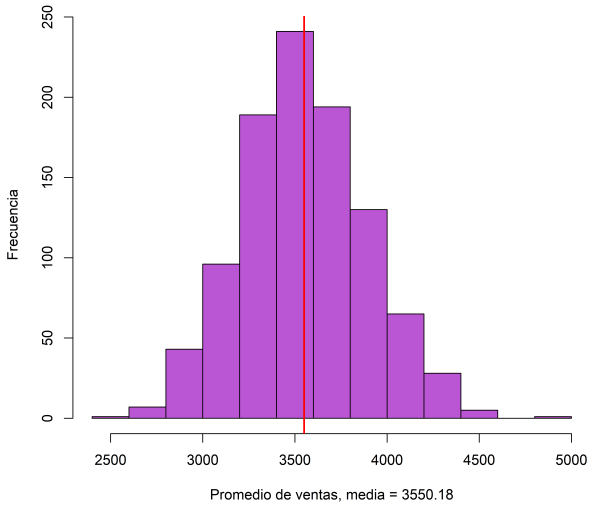
[6] The R Foundation. The R Project for Statistical Computing. https://www.r-project.org/, 2020.



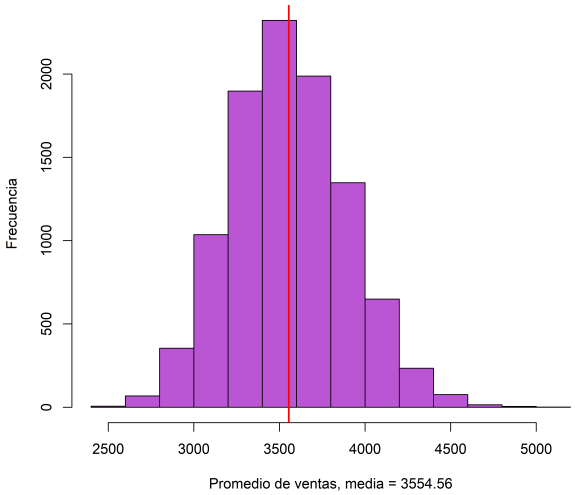

(a) $n = 10$




(b) $n = 100$

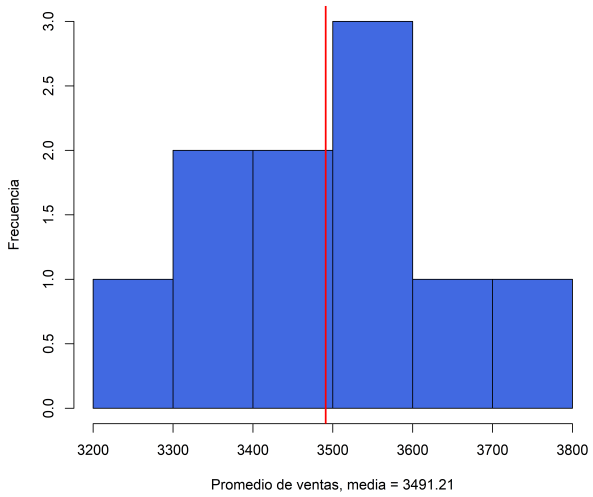



(c) $n = 1,000$

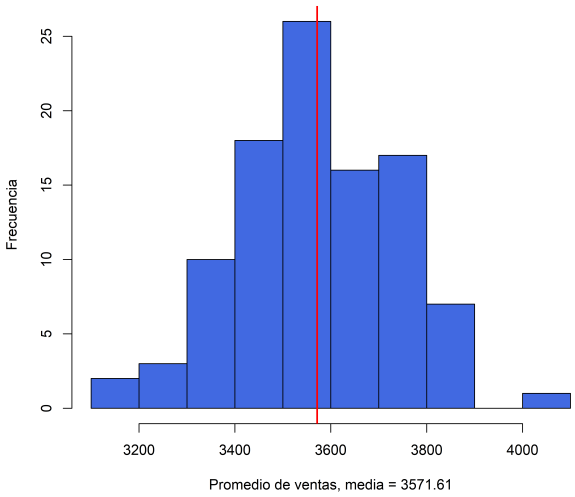



(d) $n = 10,000$

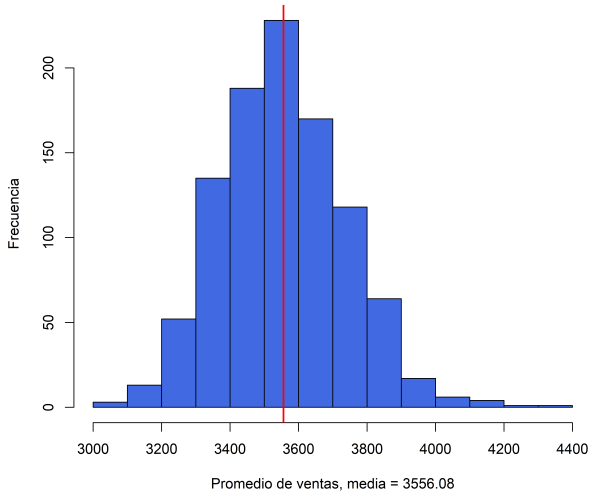
Figura 2: Resultados de la experimentación con $n$ muestras de tamaño 30



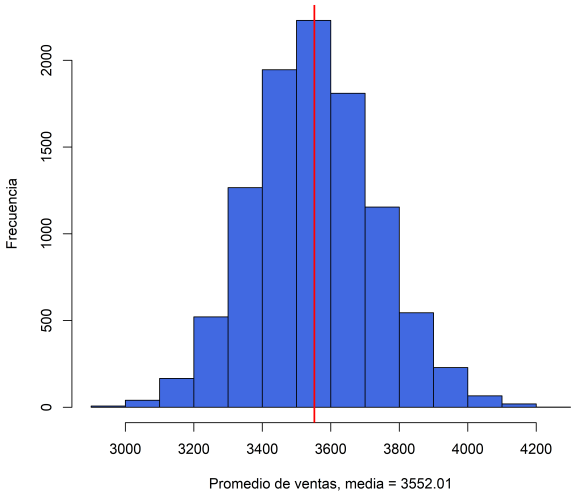

(a) $n = 10$




(b) $n = 100$




(c) $n = 1,000$




(d) $n = 10,000$

Figura 3: Resultados de la experimentación con $n$ muestras de tamaño 100